Chapter 23: Elements of Research Design

Class 25, 5/11/09 W

Slide 1 Chapter 23: Elements of Research Design

NOTES:

HW 16 due Tues 5/12/09 Noon

Submit as Myname-HW16.doc (or *.rtf)

- Read Chapter 14 Multifactor studies without replication
- For Weds read Chapter 23: Elements of Research Design
- For Monday Chapters 18-19: Comparisons of Proportions or Odds
- Final Class: Weds May 13 Research designs Designs
- Class schedule May 6 (Nesting and Experimental Designs), May 11 (Overview of generalized linear models) Exptl design May 13 W Last class
- Wimba Sessions: new times: Monday night 8 pm-9
- Homework 16: Due Tuesday 5/12/09 Noon
- Final Exam 5/22/09 Friday 8-11 am. This is the official time
  - Or 5/19/09 Tuesday 8-11 am. I'll find a room

Slide 2 HW 16 due Tues 5/12/09 Noon

NOTES:

Display 23.4

Checklist of tasks involved in the design of a study

- ☐ 1. State the objective. *What is the question of interest?*
- ☐ 2. Determine the scope of inference.
  - *Will this be a randomized experiment or an observational study?*
  - *What experimental or sampling units will be used?*
  - *What are the populations of interest?*
- ☐ 3. Understand the system under study.
- ☐ 4. Decide how to measure a response.
- ☐ 5. List factors that can affect the response.
  - Design factors
    - Factors to vary (treatments & controls)
    - Factors to fix
  - Confounding factors
    - Factors to control by design (blocking)
    - Factors to control by analysis (covariates)
    - Factors to control by randomization
- ☐ 6. Plan the conduct of the experiment (time line).
- ☐ 7. Outline the statistical analysis.
- ☐ 8. Determine the sample size ← Attempt this

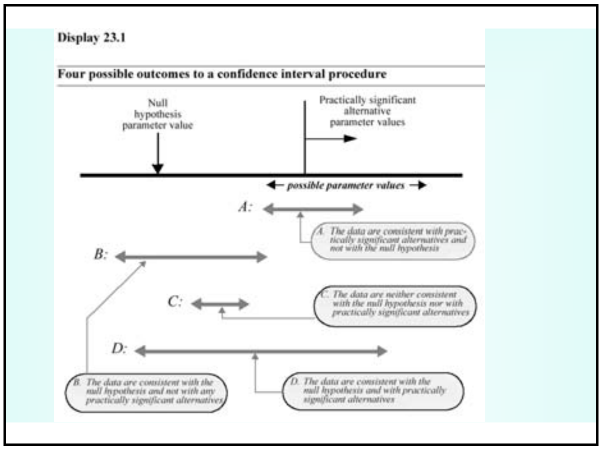
last
ork (16),
is due
ay 5/12
moved
5/11)

Slide 3

NOTES:

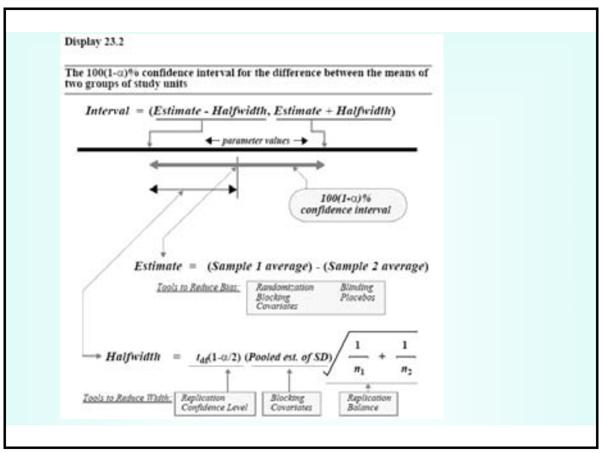
Slide 4 Elements of Research Design

NOTES:



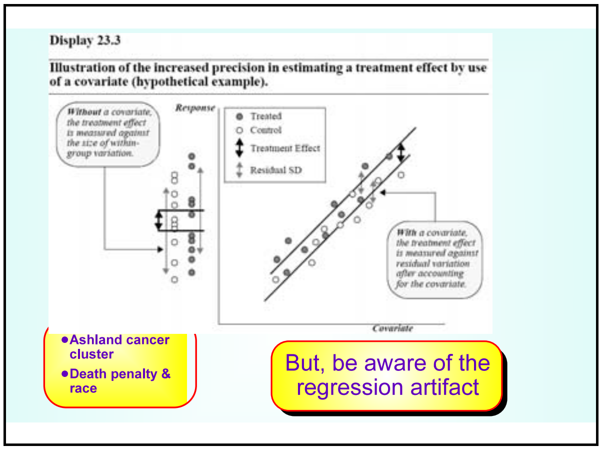
Slide 5

NOTES:



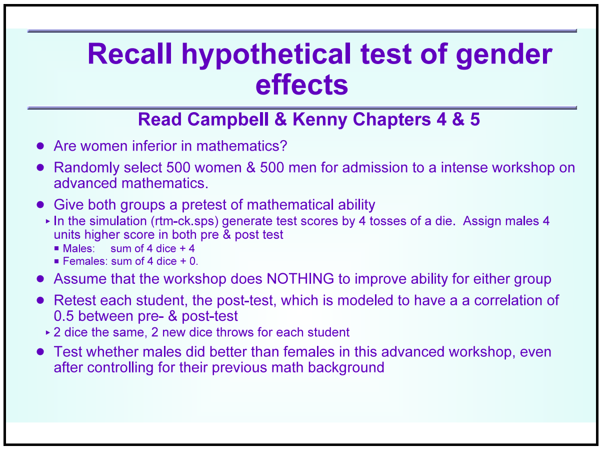
Slide 6

NOTES:



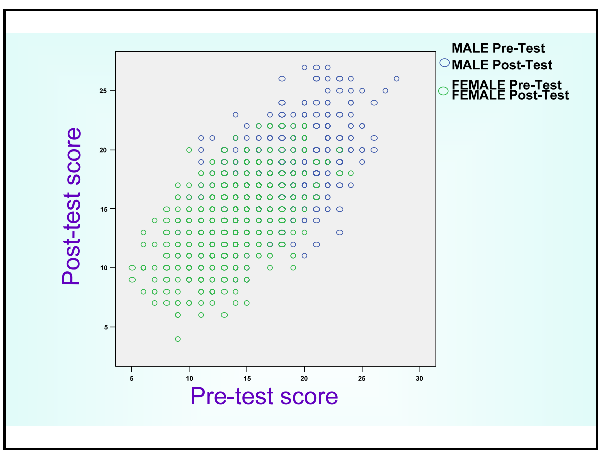
Slide 7

NOTES:



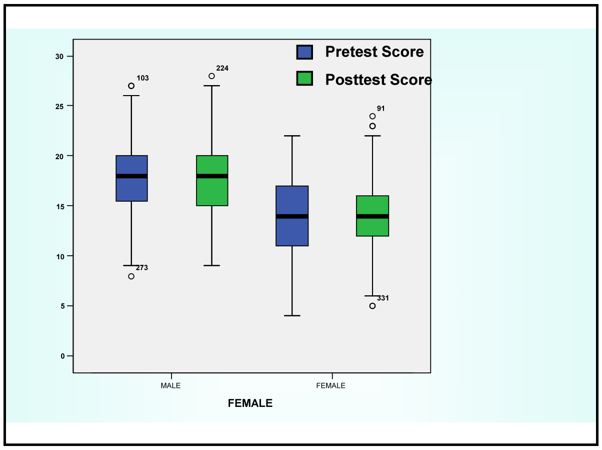
Slide 8 Recall hypothetical test of gender effects

NOTES:



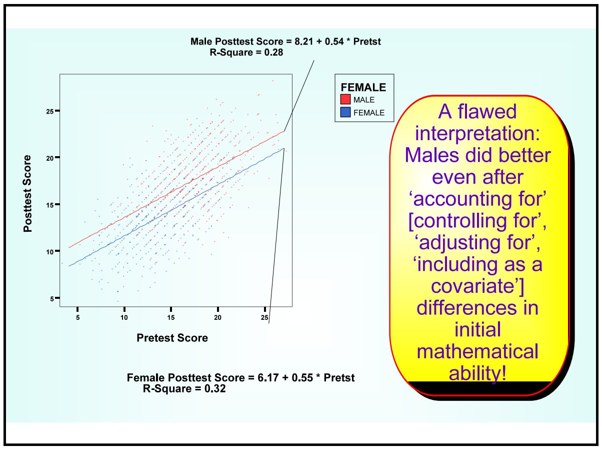
Slide 9

NOTES:



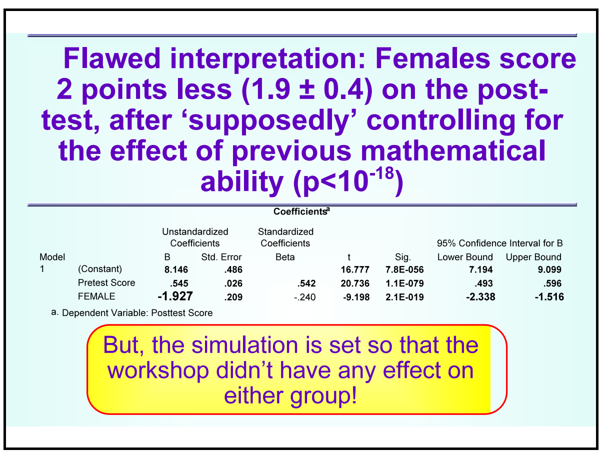
Slide 10

NOTES:



Slide 11

NOTES:



Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	8.146	.486		16.777	7.8E-056	7.194	9.099
	Pretest Score	.545	.026	.542	20.736	1.1E-079	.493	.596
	FEMALE	-1.927	.209	-.240	-9.198	2.1E-019	-2.338	-1.516

a. Dependent Variable: Posttest Score

Slide 12 Flawed interpretation: Females score 2 points less (1.9 ± 0.4) on the post-test, after 'supposedly' controlling for the effect of previous mathematical ability (p<10-18)

NOTES:

Simpson's paradox and the need to analyze data on the appropriate scale (errors due to aggregated data)

Slide 13 Simpson's paradox and the need to analyze data on the appropriate scale (errors due to aggregated data)

NOTES:

Covariates, the ecological fallacy & Simpson's paradox

- The regression artifact, improperly accounting for a covariate
  - Campbell & Kenny
  - Background effects not properly accounted for
- Simpson's Paradox & the Ecological Fallacy
  - With large scale aggregated (grouped) data, factor A may be positively associated with factor B but at smaller scales in groupings, space, or time, the factor A may really be negatively associated with factor B
  - Inferring individual responses from aggregate variables
  - This is a key error, largely ignored or unknown to analysts, in the analysis of environmental data

Slide 14 Covariates, the ecological fallacy & Simpson's paradox

NOTES:

Simpson's Paradox: failure to include covariates

http://plato.stanford.edu/entries/paradox-simpson/

At UC Berkeley, 13 males & 13 females applied for staff positions: **7/13** males hired but only **6/13** females hired

1.2 What is Simpson's Paradox?: A Diagnosis

For some whole numbers we may have:

$a/b < A/B$,

$c/d < C/D$, and

$(a + c)/(b + d) > (A + C)/(B + D)$.

Suppose that a University is trying to discriminate in favour of women when hiring staff. It advertises positions in the Department of History and in the Department of Geography, and only those departments. Five men apply for the positions in History and one is hired, and eight women apply and two are hired. The success rate for men is twenty percent, and the success rate for women is twenty-five percent. The History Department has favoured women over men. In the Geography Department eight men apply and six are hired, and five women apply and four are hired. The success rate for men is seventy-five percent and for women it is eighty percent. The Geography Department has favoured women over men. Yet across the University as a whole 13 men and 13 women applied for jobs, and 7 men and 6 women were hired. The success rate for male applicants is greater than the success rate for female applicants.

	Men		Women
History	1/5	<	2/8
Geography	6/8	<	4/5
University	7/13	>	6/13

Bickel et al. 1975 Sex bias in graduate admissions: data from Berkeley. Science

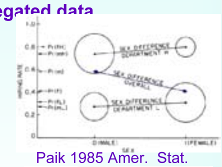
Slide 15 Simpson's Paradox: failure to include covariates

NOTES:

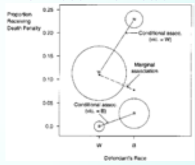
Simpson's paradox

Analyzing aggregated data

- Examples:
  - Berkeley graduate admissions: P. J. Bickel, E. A. Hammel and J. W. O'Connell (1975), "Sex bias in graduate admissions: data from Berkeley", Science 187: 398-404.
  - Agresti's death penalty case study
- An association between a pair of variables can consistently be inverted in each subpopulation of a population when the population is partitioned. e.g., a medical treatment can be associated with a higher recovery rate for treated patients compared with the recovery rate for untreated patients; yet, treated male patients and treated female patients can each have lower recovery rates when compared with untreated male patients and untreated female patients.



Paik 1985 Amer. Stat.



Slide 16 Simpson's paradox

NOTES:

Berkeley Gender discrimination

http://www.uvm.edu/~dhowell/lies4thedition/Classfolder/Simpson.html

Major Depart.	N Male Applied	N Male Admitted	% Male Admitted	N Female Applied	N Female Admitted	%Female Admitted	Female Odds Ratio
A	825	512	0.62	108	89	0.82	2.86
B	560	353	0.63	25	17	0.68	1.25
C	325	120	0.37	593	202	0.34	0.88
D	417	138	0.33	375	202	0.54	2.36
E	191	53	0.28	393	94	0.24	0.82
F	373	22	0.06	341	24	0.07	1.20
Sum	2691	1198	0.44	1835	628	0.34	0.65

Bickel et al. 1975 Sex bias in graduate admissions: data from Berkeley. Science

Slide 17 Berkeley Gender discrimination

NOTES:

Simpson's paradox & magazine subscriptions

Wagner 1982 Amer Stat.

Table 1. Expiring Subscriptions, Renewals, and Renewal Rates, by Month and Subscription Category

	Source of Current Subscription					
Month	Gift	Previous Renewal	Direct Mail	Subscription Service	Catalog Agent	Overall
January						
Total	3,594	18,364	2,986	20,862	149	45,955
Renewals	2,918	14,488	1,783	4,343	13	23,545
Rate	.812	.789	.597	.208	.087	.512
February						
Total	884	5,140	2,224	864	45	9,157
Renewals	704	3,907	1,134	122	2	5,869
Rate	.796	.760	.510	.141	.044	.641

Jan rate > Feb rate in each subcategory

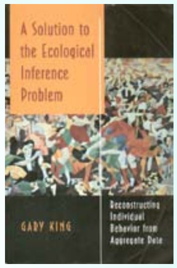
Slide 18 Simpson's paradox & magazine subscriptions

NOTES:

The ecological fallacy

Simpson's paradox & the ecological fallacy

- Ecological fallacy
  - [also called Ecological inference problem]
  - Error in predicting individual behavior from aggregated data. Introduced by Robinson (1950)
  - A solution proposed by Harvard's Gary King (1997).
- Errors can often result from inferring individual behavior from aggregated data.



Slide 19 The ecological fallacy

NOTES:

Need to control for race of victim

An example of Simpson's paradox

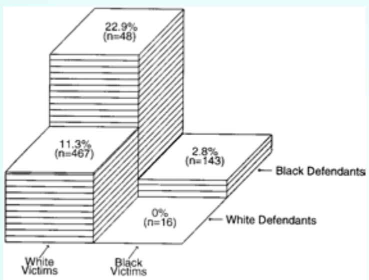
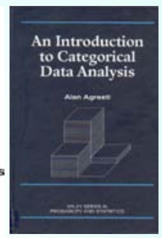


Figure 3.1 Percent receiving death penalty.



Slide 20 Need to control for race of victim

NOTES:

Is there really a racial bias in Florida death penalty cases?

White defendants are MORE likely to get the death penalty than black defendants!:
11% to 7.9%

Victim's Race	Defendant's Race	Death Penalty Yes	Death Penalty No	Total	% Yes
White	White	53	414	467	11.3%
	Black	11	37	48	22.9%
Black	White	0	16	16	0.0%
	Black	4	139	143	2.8%
Total	White	53	430	483	11.0%
	Black	15	176	191	7.9%

Agresti312deathpenalty.sav

Slide 21 Is there really a racial bias in Florida death penalty cases?

NOTES:

Death penalty conviction 'appears' independent of defendant's race

p=0.142 (1-tailed) if race of victim not considered

Defendant Race * Death Penalty Crosstabulation

			Death Penalty: Yes	Death Penalty: No	Total
Defendant Race	**White**	Count	**53**	**430**	**483**
		% within Defendant Race	11.0%	89.0%	100.0%
	Black	Count	15	176	191
		% within Defendant Race	7.9%	92.1%	100.0%
Total		Count	68	606	674
		% within Defendant Race	10.1%	89.9%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	**1.469**[b]	1	.226		
Continuity Correction[a]	**1.145**	1	.285		
Likelihood Ratio	**1.536**	1	.215		
Fisher's Exact Test				.258	.142
Linear-by-Linear Association	1.466	1	.226		
N of Valid Cases	674				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 19. 27.

Slide 22 Death penalty conviction 'appears' independent of defendant's race

NOTES:

Must include race of victim as covariate

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			**.412**
ln(Estimate)			-.887
Std. Error of ln(Estimate)			.371
Asymp. Sig. (2-sided)			.017
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	**.199**
		Upper Bound	**.852**
	ln(Common Odds Ratio)	Lower Bound	-1.614
		Upper Bound	-.160

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

Calculate inverse of odds ratios or transpose a col or row:
([0.852 0.412 0.199]).^-1
ans = 1.1737 2.4272 5.0251

The odds of a black defendant getting death penalty are 2.4 times higher than a white defendant when victim's race is considered (p=0.02, 95% CI 1.17 to 5.03)

Slide 23 Must include race of victim as covariate

NOTES:

Simpson's paradox

Driven by strong association between victim's & defendant's race, Agresti (1996, Fig 3.2)

11.3% if victim white

11.0% of all white defendants given the death penalty

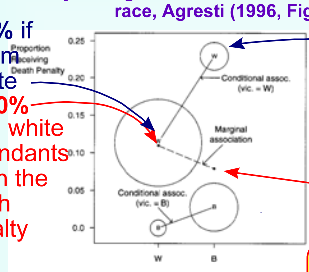


Figure 3.2 Proportion receiving death penalty by defendant's race, controlling and ignoring victim's race

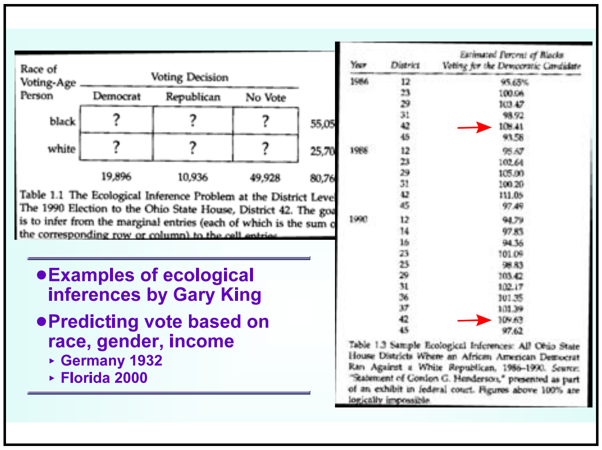
22.9% If victim white, black odds 2.3 times white (95% CI: 1.1 to 4.8)

7.9% of black defendants

If victim's race not considered, little difference between blacks & whites

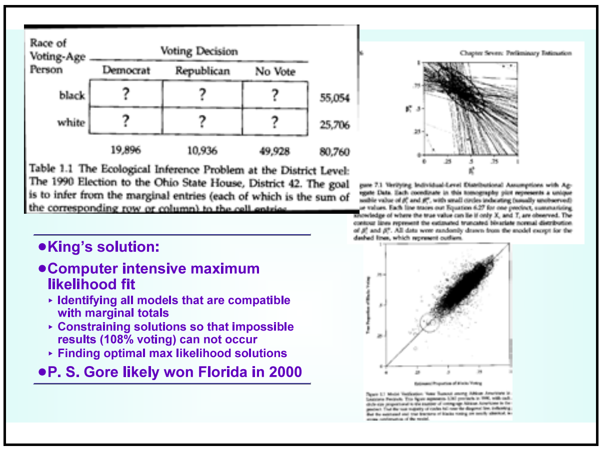
Slide 24 Simpson's paradox

NOTES:



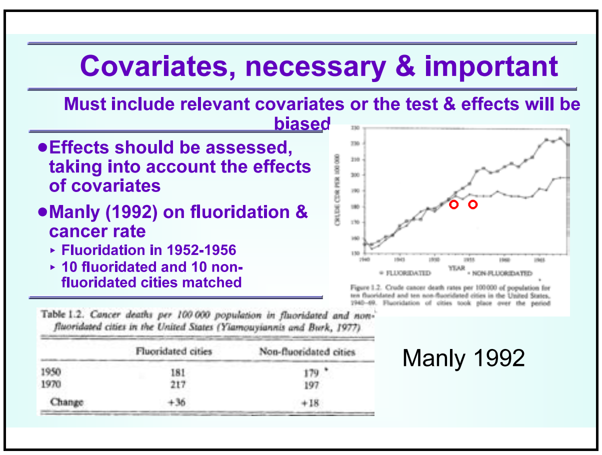
Slide 25

NOTES:



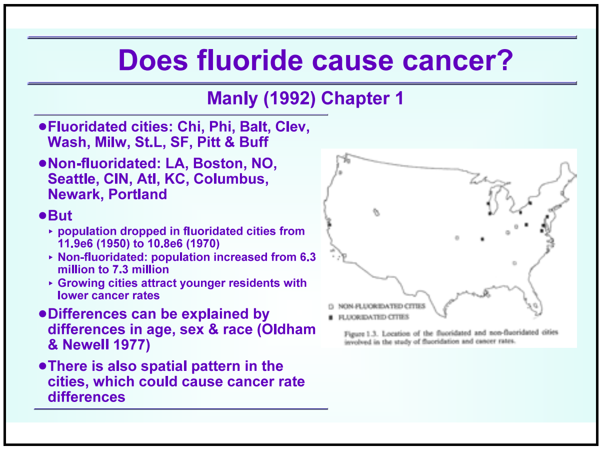
Slide 26

NOTES:



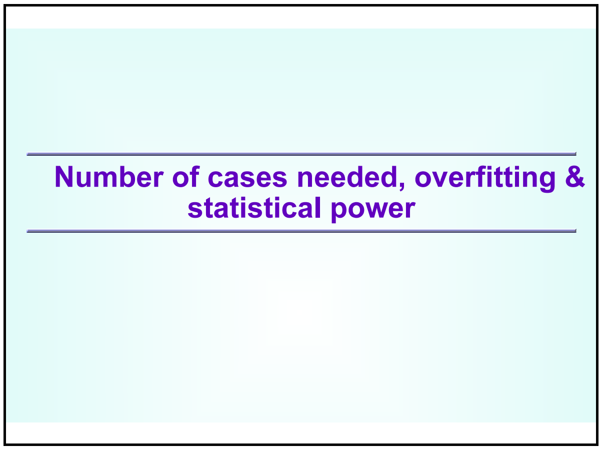
Slide 27 Covariates, necessary & important

NOTES:



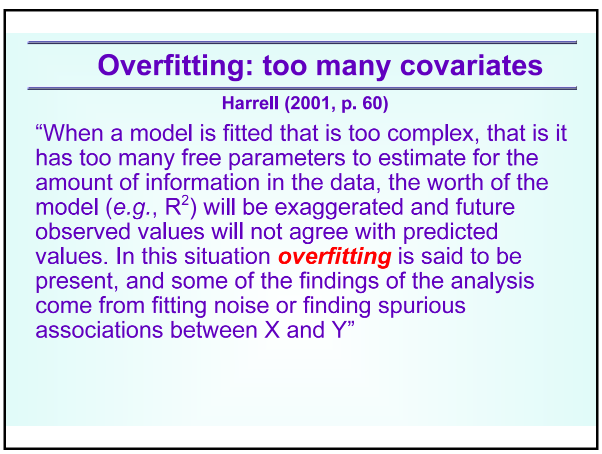
Slide 28 Does fluoride cause cancer?

NOTES:



Slide 29 Number of cases needed, overfitting & statistical power

NOTES:



Slide 30 Overfitting: too many covariates

NOTES:

Number of cases needed for regression (1 of 2)

Harrell (2001, p. 61)

- Number of predictors should be less than m/10 or m/20 where m is the limiting sample size shown below
- Candidate variables must include all variables screened for association with response, including nonlinear terms and interactions

TABLE 4.1: Limiting Sample Sizes for Various Response Variables

Type of Response Variable	Limiting Sample Size m
Continuous	n (total sample size)
Binary	$\min(n_1, n_2)$ [c]
Ordinal (k categories)	$n - \frac{1}{n^2}\sum_{i=1}^{k} n_i^3$ [d]
Failure (survival) time	number of failures [e]

Slide 31 Number of cases needed for regression (1 of 2)

NOTES:

Number of cases for regression (2 of 2)

Tabachnik & Fidell (2001, p 117)

- For multiple regression (from Green 1991)
  - $N \geq 50 + 8m$, where m is the number of explanatory variables, for testing R^2, and
  - $N \geq 104 + m$ for individual predictors
  - A higher case to explanatory variable ratio is needed when
    - Effect sizes are small
    - Data are skewed
    - Measurement error is expected in explanatory variables
  - Automated selection procedures (statistical regression)
    - Cases > 40 * explanatory variables
  - Green's more precise rule
    - $N \geq (8 / f^2) + (m-1)$, where f^2 = 0.01, 0.15, and 0.35 for small, medium and large effect sizes.
    - $f^2 = R^2 / (1-R^2)$, where R^2 is the expected squared multiple correlation coefficient

Slide 32 Number of cases for regression

(2 of 2)

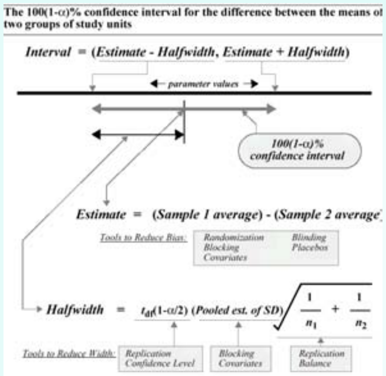
NOTES:

Power analysis

Prospective not retrospective

Slide 33 Power analysis

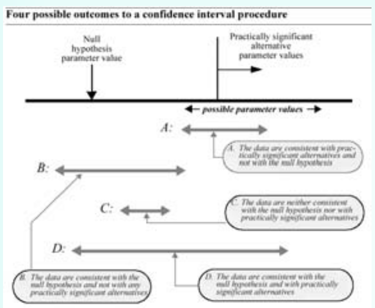
NOTES:



Slide 34

NOTES:

Neyman-Pearson CI's



Slide 35 Neyman-Pearson CI's

NOTES:

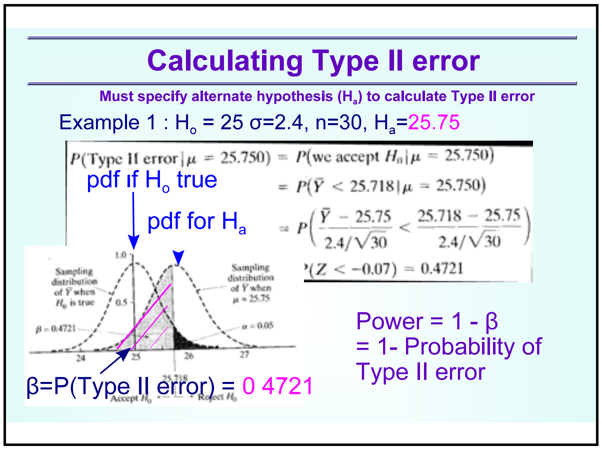
Type I and II errors

Accept H_o should be 'tentatively accept H_o'

		True State of Nature	
		H_0 is true	H_1 is true
Our Decision	Accept H_0	Correct decision	Type II error
	Reject H_0	Type I error	Correct decision

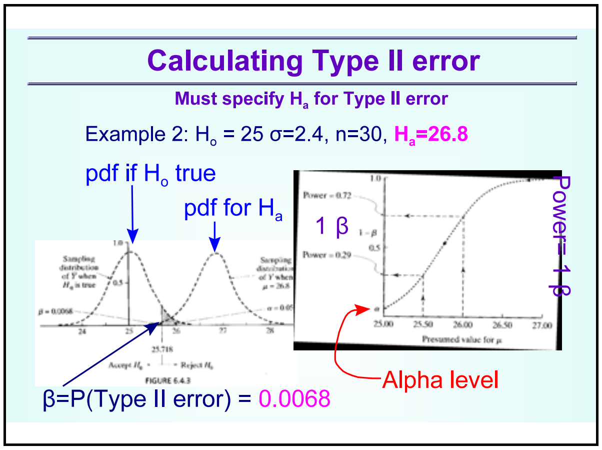
Slide 36 Type I and II errors

NOTES:



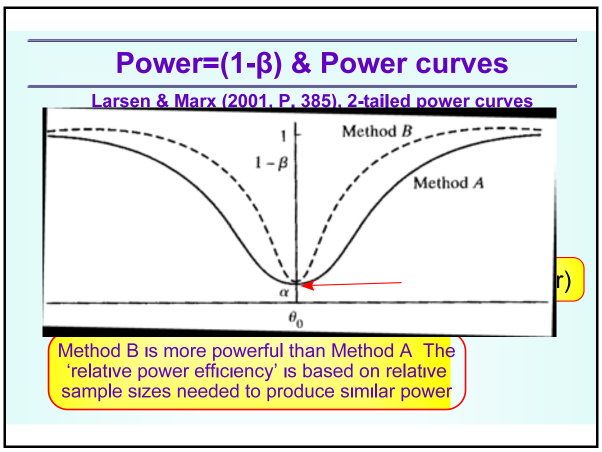
Slide 37 Calculating Type II error

NOTES:



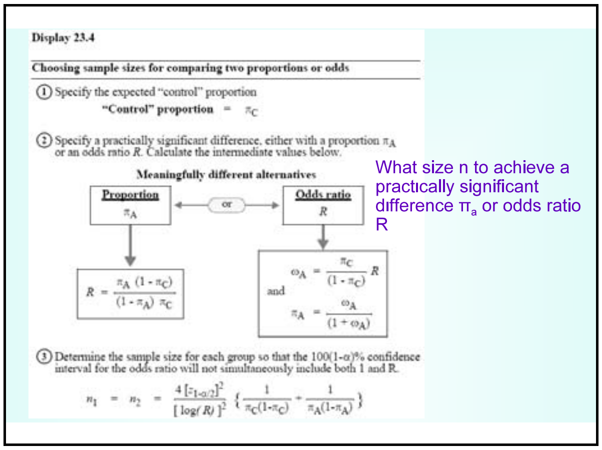
Slide 38 Calculating Type II error

NOTES:



Slide 39 Power=(1-β) & Power curves

NOTES:

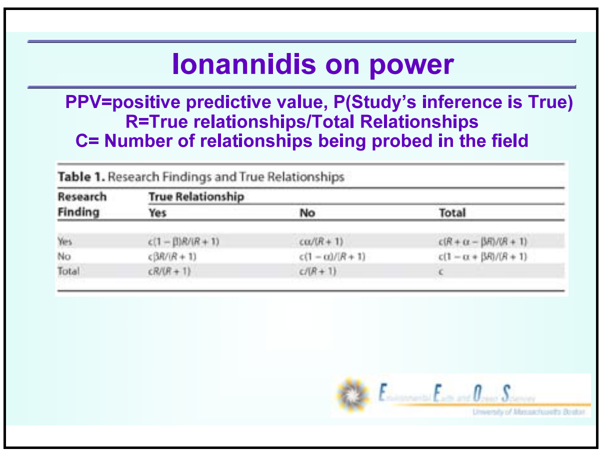


Slide 40

NOTES:

Slide 41 Ionannidis on power

NOTES:



Slide 42 Ionannidis on power

NOTES:

Ionnidas on 'bias,' should be fraud

Not the accepted meaning of bias

Bias

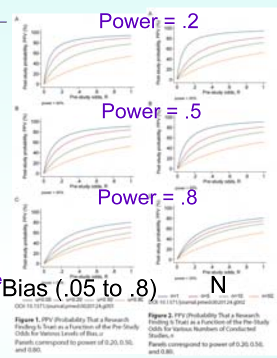
First, let us define bias as the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced. Let u be the proportion of probed analyses that would not have been "research findings," but nevertheless end up presented and reported as such, because of bias. Bias should not be confused with chance variability that causes some findings to be false by chance even though the study design, data, analysis, and presentation are perfect. Bias can entail manipulation in the analysis or reporting of findings. Selective or distorted reporting is a typical form of such bias. We may assume that u does not depend on whether a true relationship exists or not. This is not an unreasonable assumption, since typically it is impossible to know which relationships are indeed true. In the presence of bias (Table 2), one gets PPV = $([1 - \beta]R + u\beta R)/(R + \alpha - \beta R + u - u\alpha + u\beta R)$, and PPV decreases with increasing u, unless $1 - \beta \leq \alpha$, i.e., $1 - \beta \leq 0.05$ for most situations. Thus, with increasing bias, the chances that a research finding is true diminish considerably. This is shown for different levels of power and for different pre-study odds in Figure 1.

Bias in statistics: The difference between the expected value and the true value of a parameter ***cf.*****, unbiased estimator**

Slide 43 Ionnidas on 'bias,' should be fraud

NOTES:

- Corollary 1: The smaller the study's sample size, the less likely the results are to be true. Low sample size produces tests with low power (Large clinical trials more likely to produce true results)
- Corollary 2: The smaller the effect size, the less likely the result is true
- Corollary 3: The greater the number of studies, the less likely the result is to be true
- Corollary 4: The greater the 'flexibility' in analysis, the less likely the result
- Corollary 5: The greater the financial incentive, the less likely a result is to be true
- Corollary 6: The hotter the scientific field, the less likely the result is to be true



Slide 44

NOTES:

Ionnades' recommendations

- Perform studies only if the sample sizes are large enough to ensure high power
- Register the study, design and hypotheses in advance to avoid the identification of significant results that are spurious
- Design experiments and surveys to test hypotheses with high initial probabilities of being true
  - Often relationships assumed to be true in a field are not true.
  - Test established foundations of a field

Slide 45 Ionnades' recommendations

NOTES:

Retrospective power analyses

Hoenig & Heisey (2001): The abuse of power

- The dilemma of the nonrejected null hypothesis: what should we do?
- 19 applied journals, including Ecology, required *post-hoc* power calculations
- Winer *et al.* (1991) & Zar (1996) recommend post-hoc power tests
- Dayton (1998): reverse the burden of proof: How big could the effect have been and still have been missed? The no-impact null.
- Alternative recommended by Hoenig & Heisey (2001): interpret confidence intervals & discuss sample size issues

Slide 46 Retrospective power analyses

NOTES:

Retrospective power analyses

Hoenig & Heisey (2001): The abuse of power (2 of 2)

- Observed power, available in SPSS
  - Case Study 2.1 Bumpus's sparrows
- Student's test found a 0.01 inch difference but an independent samples t test found a 2-sided P value of 0.08
- UNIANOVA can estimate the observed power for this design

Slide 47 Retrospective power analyses

NOTES:

Case Study 2.1

Observed power available in GLM Univariate, but don't use!

Independent Samples Test

		t-test for Equality of Means						
							95% Confidence Interval of the Difference	
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Humerus length (in.x1000)	Equal variances assumed	-1.777	57	.081	-10.083	5.674	-21.446	1.279

Parameter Estimates

Dependent Variable: Humerus length (in.x1000)

					95% Confidence Interval		
Parameter	B	Std. Error	t	Sig.	Lower Bound	Upper Bound	Observed Power[a]
Intercept	738.000	3.619	203.920	.000	730.753	745.247	1.000
[group=1]	-10.083	5.674	-1.777	.081	-21.446	1.279	.416
[group=2]	0[b]	.	.	.	.	.	.

a. Computed using alpha = .05

b. This parameter is set to zero because it is redundant.

With the observed standard error, the probability of Type II error is 58.4% (1-Power) against an alternate hypothesis of 0.01 inch larger humerus in those that survived

Slide 48 Case Study 2.1

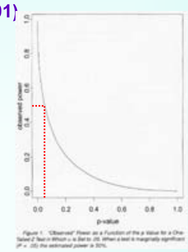
NOTES:

What's wrong with power analysis?

Hoenig & Heisey (2001)

- **Observed power is determined completely by the p value and adds nothing more**
- **If Z = alpha for a 1-tailed test, then the observed power is 0.5**

If the difference was exactly 10.083 inches, and the difference was symmetric, then there would be a 0.5 probability of rejecting the null hypothesis at α=0.05



Dependent Variable: Humerus length (in.x1000)

Parameter Estimates

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound	Partial Eta Squared	Noncent. Parameter	**Observed Power**[a]
Intercept	738.000	3.619	203.920	.000	730.753	745.247	.999	203.920	**1.000**
[group=1]	-10.083	5.674	-1.777	.081	-21.446	1.279	.052	1.777	**.416**
[group=2]	0[b]	.	.	.	.	.	.	.	.

a. Computed using alpha = .05

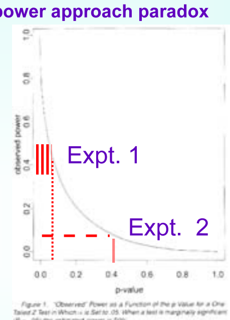
Slide 49 What's wrong with power analysis?

NOTES:

What's wrong with power analysis?

Hoenig & Heisey (2001): The power approach paradox

- **Many authors argue**
  - that the higher the observed power, the greater the evidence in favor of the null hypothesis.
  - Conversely, low power offers only weak support for the null hypothesis
  - Hoenig & Heisey: "This is easily shown to be nonsense."
- **Imagine 2 experiments.**
  - In experiment 1, the p value is 0.08 which offers only weak evidence against the null. The power is 0.42
  - In experiment 2, the p value is 0.4 which offers much stronger evidence that the null hypothesis is true against similar alternative hypotheses.
  - However, the observed power in the 2nd experiment is only 0.1, which would indicate weaker evidence in favor of the truth of the null hypothesis
- **"Higher observed power does not imply stronger evidence for a null hypothesis that is not rejected."**



Slide 50 What's wrong with power analysis?

NOTES:

Detectable effect size: also bad

Hoenig & Heisey (2001)

- Those that argue for post hoc power analysis require an answer to the question, "What is the effect size required to achieve a power of 90%?"
  - This would be the detectable effect size
- The closer the detectable effect size is to zero, the stronger the evidence is taken to be **for** the null hypothesis
- Imagine two experiments with the same effect size & same sample size, but $Z_1 > Z_2$, $p_1 < p_2$ which implies $\sigma_1 < \sigma_2$
- The detectable effect size will be smaller in the 1st experiment
- ...leading to the nonsensical conclusion that the 1st experiment with the lower p value (e.g, 0.06) provides stronger evidence for the null hypothesis being true than the 2nd experiment with the higher p value (e.g, 0.4)

Slide 51 Detectable effect size: also bad

NOTES:

Alternatives to post-hoc power analyses

Hoenig & Heisey (2001)

- Use confidence intervals: once the confidence interval is calculated, power analysis provides no further insights.
- “We believe that the central focus of data analysis should be to find which parameter values are supported by the data and which are not.”
- Bayesian posterior probabilities offer a solution to these problems
- Statistics classes should place more emphasis on confidence intervals and less on hypothesis testing and p values
  - Researchers interpret frequentist CI’s as Bayesian credibility regions: so what?

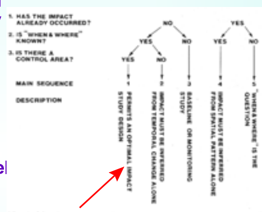
Slide 52 Alternatives to post-hoc power analyses

NOTES:

BACI designs

Before-After-Control-Impact design

- **Described by Green (1979)**
- **Green argued that one could use an Optimal impact study design: use a 2-way ANOVA with the interaction effect being the key test statistic**
- **Hurlbert (1984) attacked this view**
- **Paul Murtaugh has a recent critique of recent BACI model (fail to assess serial correlation effects)**



The decision key to the "main sequence" categories of environmental studies.

Slide 53 BACI designs

NOTES:

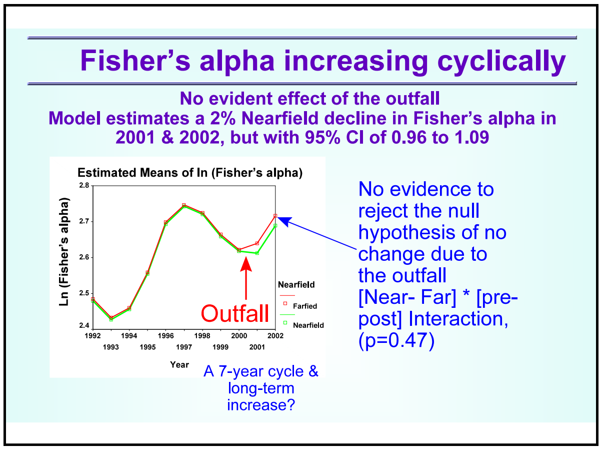
BACI designs criticized

If 1 treatment & 1 control area

- Green (1979) use a site x time interaction term
- Hurlbert (1984) assumes that 2 sites remain parallel
- Stewart-Oaten & Murdoch: Measure the differences between sites multiple times before and after impact
  - A form of repeated measures design
- Murtaugh (Ecology 2000, 2002):
  - BACI designs ignore serial correlation
  - Murtaugh: P (Type I error) ≈20% with real data with positive serial correlation
  - Adjusting for serial correlation produces tests with little power
  - Solution: just plot the data and avoid significance tests
- Murtaugh (2003): No p values are better than incorrect ones. Don’t use inferential statistics if the design is bad, just report the data

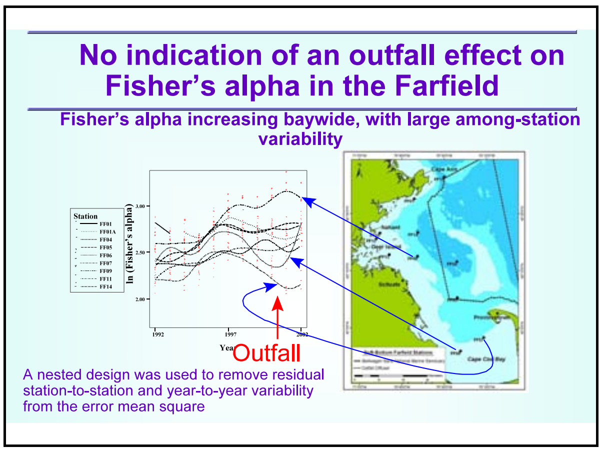
Slide 54 BACI designs criticized

NOTES:



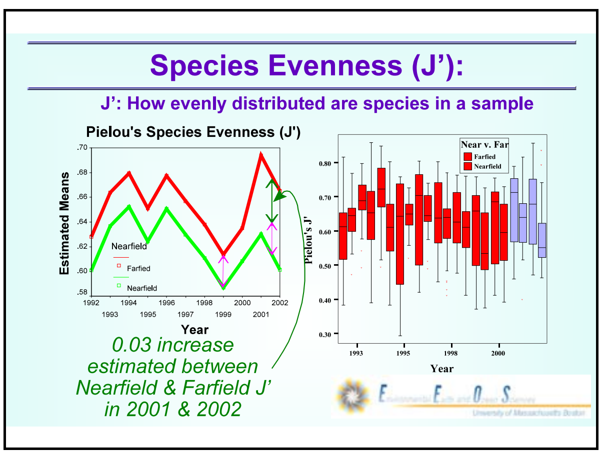
Slide 55 Fisher's alpha increasing cyclically

NOTES:



Slide 56 No indication of an outfall effect on Fisher's alpha in the Farfield

NOTES:

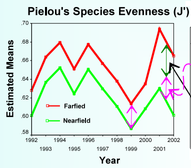


Slide 57 Species Evenness (J'):

NOTES:

Pielou's Evenness

A 5% increase in Farfield relative to Nearfield in 2001 & 2002, Effect, indicated with green arrowheads, tested with the Near -Far x Pre-Post Outfall Interaction term



Tests of Between-Subjects Effects

Dependent Variable: Pielou's J'

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	95.049	1	95.049	2286.	
	Error	1.654	39.796	.042		
NEAR vs Far	Hypothesis	.114	1	.114	2.9	.102
	Error	1.630	40.108	.041		
PRE vs Post	Hypothesis	.013	1	.013	2.1	.146
	Error	3.645	597	.006		
NEAR * PRE	Hypothesis	.031	1	.031	5.1	.024
	Error	3.645	597	.006		
PRE(YEAR)	Hypothesis	.256	9	.028	4.7	< E-05
	Error	3.645	597	.006		
NEAR(STATION)	Hypothesis	2.701	34	.079	13.0	< E-6
	Error	3.645	597	.006[d]		

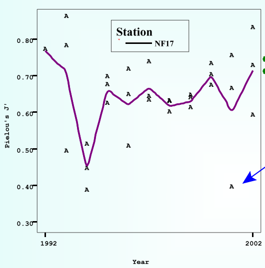
My model is a 2-factor mixed model
Nested ANOVA: Pre vs. Post
nested within Years, Near vs. Far
nested within station effects

Slide 58 Pielou's Evenness

NOTES:

High variability in Pielou's Species Evenness (J')

Especially at sandy NF-17



Significant effect size from ANOVA

1579 *Molgula manhattensis* in 1 sample (64% of the Total)

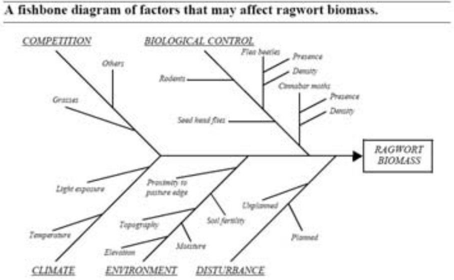
Slide 59 High variability in Pielou's Species Evenness (J')

NOTES:

Ragwort example

Display 23.6

A fishbone diagram of factors that may affect ragwort biomass.



Slide 60 Ragwort example

NOTES:

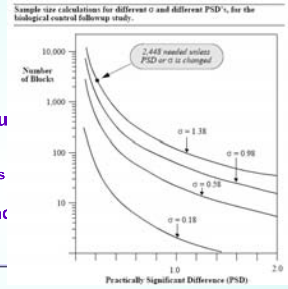
Ragwort example

2,451 blocks required: an impossibility

- Solutions (Sleuth p 687)
- Decrease the level of confidence
- Increase the size of the practical significance
- Consider a repeated measu crosssover design
  - Often not an option
  - See Appendix for crossover desi issues
- Reduce the residual varianc
  - Blocking
  - Adding covariates

Display 23.7

Sample size calculations for different σ and different PSD's, for the biological control followup study.



Slide 61 Ragwort example

NOTES:

TABLE 1. Potential sources of confusion in an experiment and means for minimizing their effect.

Source of confusion	Features of an experimental design that reduce or eliminate confusion
1. Temporal change	Control treatments
2. Procedure effects	Control treatments
3. Experimenter bias	Randomized assignment of experimental units to treatments Randomization in conduct of other procedures "Blind" procedures*
4. Experimenter-generated variability (random error)	Replication of treatments
5. Initial or inherent variability among experimental units	Replication of treatments Interspersion of treatments Concomitant observations
6. Nondemonic intrusion†	Replication of treatments Interspersion of treatments
7. Demonic intrusion	Eternal vigilance, exorcism, human sacrifices, etc.

* Usually employed only where measurement involves a large subjective element.
† Nondemonic intrusion is defined as the impingement of chance events on an experiment in progress.

Hurlbert (1984) on exper mental des gn

Slide 62

NOTES:

Random or systematic?

Hurlbert (1984) & Underwood argue for systematic sampling designs to avoid aggregation which might occur by chance

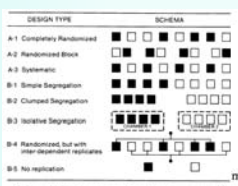


FIG. 1. Schematic representation of various acceptable modes (A) of interspersing the replicates (boxes) of two treatments (shaded, unshaded) and various ways (B) in which the principle of interspersion can be violated.

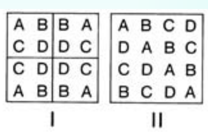
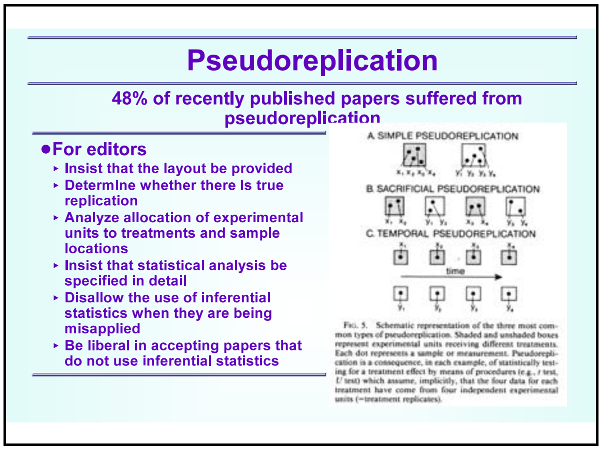


FIG. 3. Examples of segregated arrangements of four treatments, each replicated four times, that can result from use of restricted randomization procedures: (I) randomized block design, (II) Latin square design.

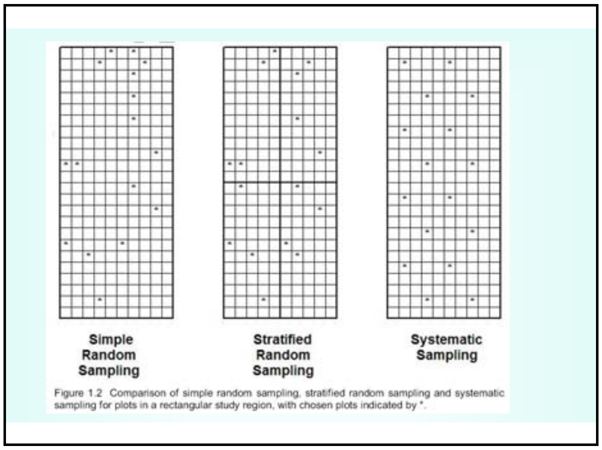
Slide 63 Random or systematic?

NOTES:



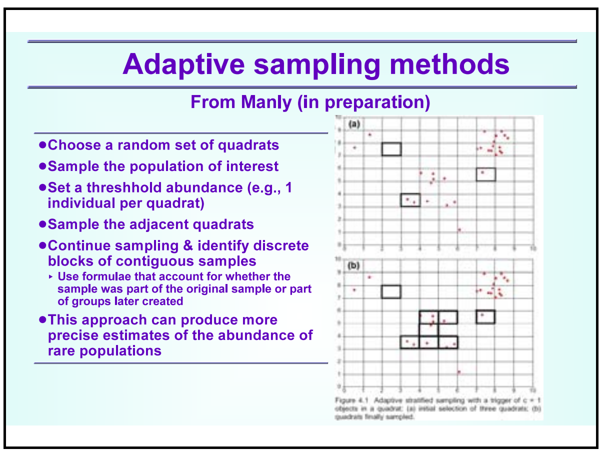
Slide 64 Pseudoreplication

NOTES:



Slide 65

NOTES:



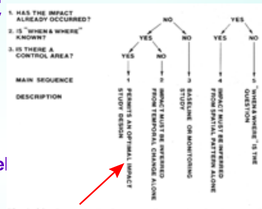
Slide 66 Adaptive sampling methods

NOTES:

BACI designs

Before-After-Control-Impact design

- **Described by Green (1979)**
- **Green argued that one could use an Optimal impact study design: use a 2-way ANOVA with the interaction effect being the key test statistic**
- **Hurlbert (1984) attacked this view**
- **Paul Murtaugh has a recent critique of recent BACI model (fail to assess serial correlation effects)**



The decision key to the "main sequence" categories of environmental studies.

Slide 67 BACI designs

NOTES:

BACI designs criticized

If 1 treatment & 1 control area

- Green (1979) use a site x time interaction term
- Hurlbert (1984) assumes that 2 sites remain parallel
- Stewart-Oaten & Murdoch: Measure the differences between sites multiple times before and after impact
  - A form of repeated measures design
- Murtaugh (Ecology 2000, 2002):
  - BACI designs ignore serial correlation
  - Murtaugh: P (Type I error) ≈20% with real data with positive serial correlation
  - Adjusting for serial correlation produces tests with little power
  - Solution: just plot the data and avoid significance tests
- Murtaugh (2003): No p values are better than incorrect ones. Don't use inferential statistics if the design is bad, just report the data

Slide 68 BACI designs criticized

NOTES:

Ecology, 67(4), 1986, pp. 929–940

ENVIRONMENTAL IMPACT ASSESSMENT: "PSEUDOREPLICATION" IN TIME?[1]

Allan Stewart-Oaten and William W. Murdoch
Department of Biological Sciences, University of California, Santa Barbara, California 93106 USA

AND

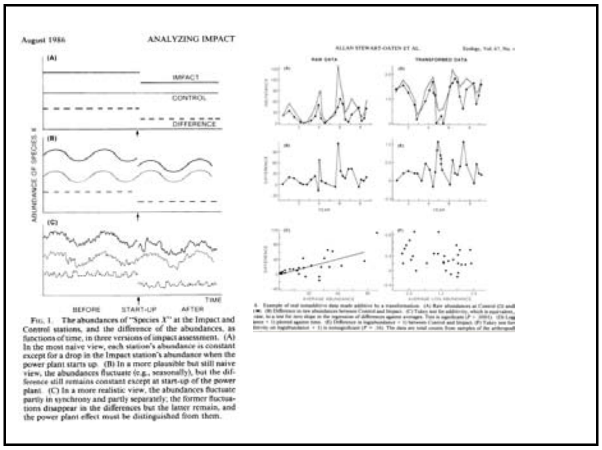
Keith R. Parker
Marine Review Committee, 531 Encinitas Boulevard, Encinitas, California 92024 USA

Abstract. A recent monograph by Hurlbert raised several problems concerning the appropriate design of sampling programs to assess the impact upon the abundance of biological populations of, for example, the discharge of effluents into an aquatic ecosystem at a single point. Key to the resolution of these issues is the correct identification of the statistical parameter of interest, which is the mean of the underlying probabilistic "process" that produces the abundance, rather than the actual abundance itself. We describe an appropriate sampling scheme designed to detect the effect of the discharge upon this underlying mean. Although not guaranteed to be universally applicable, the design should meet Hurlbert's objections in many cases. Detection of the effect of the discharge is achieved by testing whether the *difference* between abundances at a control site and an impact site changes once the discharge begins. This requires taking samples, replicated in time, Before the discharge begins and After it has begun, at both the Control and Impact sites (hence this is called a BACI design). Care needs to be taken in choosing a control site so that it is sufficiently far from the discharge to be largely beyond its influence, yet close enough that it is influenced by the same range of natural phenomena (e.g., weather) that result in long-term changes in the biological populations. The design is not appropriate where local events cause populations at Control and Impact sites to have different long-term trends in abundance; however, these situations can be detected statistically. We discuss the assumptions of BACI, particularly additivity (and transformations to achieve it) and independence.

Key words: *environmental monitoring; impact assessment; independence; pollutants; power plants; replication; serial correlation; statistical transformations; statistics.*

Slide 69

NOTES:

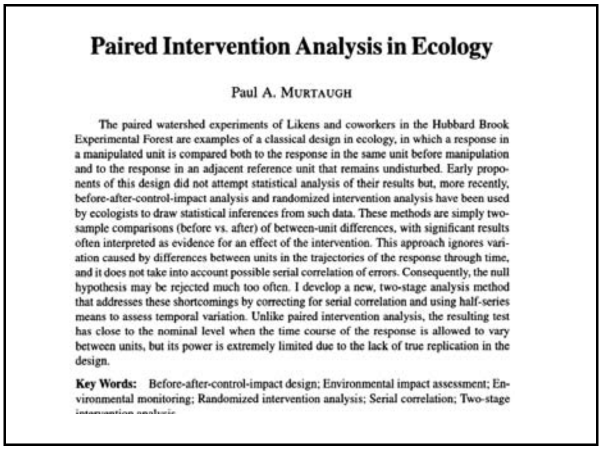


August 1986 ANALYZING IMPACT

FIG. 1. The abundances of "Species X" at the Impact and Control stations, and the difference of the abundances, as functions of time, in three versions of impact assessment. (A) In the most naive view, each station's abundance is constant except for a drop in the Impact station's abundance when the power plant starts up. (B) In a more plausible but still naive view, the abundances fluctuate (e.g., seasonally), but the difference still remains constant except at start-up of the power plant. (C) In a more realistic view, the abundances fluctuate partly in synchrony and partly separately; the former fluctuations disappear in the differences but the latter remain, and the power plant effect must be distinguished from them.

Slide 70

NOTES:



Paired Intervention Analysis in Ecology

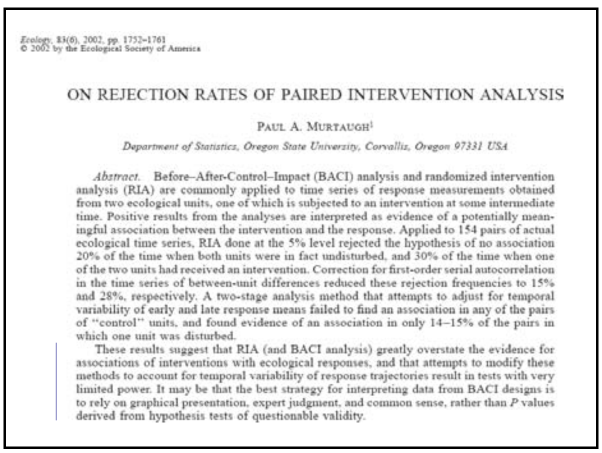
Paul A. Murtaugh

The paired watershed experiments of Likens and coworkers in the Hubbard Brook Experimental Forest are examples of a classical design in ecology, in which a response in a manipulated unit is compared both to the response in the same unit before manipulation and to the response in an adjacent reference unit that remains undisturbed. Early proponents of this design did not attempt statistical analysis of their results but, more recently, before-after-control-impact analysis and randomized intervention analysis have been used by ecologists to draw statistical inferences from such data. These methods are simply two-sample comparisons (before vs. after) of between-unit differences, with significant results often interpreted as evidence for an effect of the intervention. This approach ignores variation caused by differences between units in the trajectories of the response through time, and it does not take into account possible serial correlation of errors. Consequently, the null hypothesis may be rejected much too often. I develop a new, two-stage analysis method that addresses these shortcomings by correcting for serial correlation and using half-series means to assess temporal variation. Unlike paired intervention analysis, the resulting test has close to the nominal level when the time course of the response is allowed to vary between units, but its power is extremely limited due to the lack of true replication in the design.

Key Words: Before-after-control-impact design; Environmental impact assessment; Environmental monitoring; Randomized intervention analysis; Serial correlation; Two-stage

Slide 71

NOTES:



Ecology, 83(6), 2002, pp. 1752–1761
© 2002 by the Ecological Society of America

ON REJECTION RATES OF PAIRED INTERVENTION ANALYSIS

Paul A. Murtaugh[1]

Department of Statistics, Oregon State University, Corvallis, Oregon 97331 USA

Abstract. Before–After-Control–Impact (BACI) analysis and randomized intervention analysis (RIA) are commonly applied to time series of response measurements obtained from two ecological units, one of which is subjected to an intervention at some intermediate time. Positive results from the analyses are interpreted as evidence of a potentially meaningful association between the intervention and the response. Applied to 154 pairs of actual ecological time series, RIA done at the 5% level rejected the hypothesis of no association 20% of the time when both units were in fact undisturbed, and 30% of the time when one of the two units had received an intervention. Correction for first-order serial autocorrelation in the time series of between-unit differences reduced these rejection frequencies to 15% and 28%, respectively. A two-stage analysis method that attempts to adjust for temporal variability of early and late response means failed to find an association in any of the pairs of "control" units, and found evidence of an association in only 14–15% of the pairs in which one unit was disturbed.

These results suggest that RIA (and BACI analysis) greatly overstate the evidence for associations of interventions with ecological responses, and that attempts to modify these methods to account for temporal variability of response trajectories result in tests with very limited power. It may be that the best strategy for interpreting data from BACI designs is to rely on graphical presentation, expert judgment, and common sense, rather than P values derived from hypothesis tests of questionable validity.

Slide 72

NOTES:

Comments

Ecology, 84(10), 2003, pp. 2795–2799

ON REJECTION RATES OF PAIRED INTERVENTION ANALYSIS: COMMENT

Allan Stewart-Oaten[1]

Murtaugh (2000, 2002) claims the Before–After, Control–Impact (BACI) approach to assessment of long-term local effects of a planned environmental alteration (such as a development) ignores serial correlation and assumes the Control and Impact values would have "parallel trajectories" in the absence of an impact.

Ecology, 84(10), 2003, pp. 2799–2802

ON REJECTION RATES OF PAIRED INTERVENTION ANALYSIS: REPLY

Paul A. Murtaugh[1]

Stewart-Oaten (2003) criticizes my paper on paired intervention analysis (Murtaugh 2002) on several grounds. By "paired intervention analysis" I mean before–after, control–impact (BACI) analysis and randomized intervention analysis (RIA) applied to data from a single pair of ecological units.

The problem of serial correlation

Increasing numbers of authors are looking for, and attempting to adjust for, serial correlation (Hewitt et al. 2001, Levin and Tolimieri 2001, Rumbold et al. 2001, Zimmer et al. 2001), but many others continue to overlook the problem (Basset et al. 2001, Guidetti 2001, Guillemette and Larsen 2002, Roman et al. 2002, Rybczyk et al. 2002). They may have good reason: it

Does that mean that unreplicated *ecosystem-level* manipulations are without merit? Of course not. Would the studies of Likens et al. (1970) on a single pair of watersheds be more compelling if they had been accompanied by BACI-derived P values? Would their results have been less compelling if *three* pairs of watersheds had been used, and an analysis correctly based on this level of replication yielded $P = 0.10$? In my opinion, this sort of slavish devotion to P values (and, yes, confidence intervals) gets in the way of good science.

Stewart-Oaten (2003) views my skepticism about BACI analyses as an "abdication of responsibility," an abandonment of the objectivity that we must bring to scientific investigations. I would respond that *no* P values are better than incorrect ones. As a statistician, I could not agree more that "properly carried out and explained, [statistical inference] can be thought of as a systematized, objective form of common sense." Improperly carried out, statistical inference can be misleading, distracting, and detrimental to the progress of science.

Slide 73
NOTES: